Supplementary Information (SI) for:

Correlating Measures of Hierarchical Structures in Artificial Neural Networks with Their Performance

Zhuoying Xu,[1, a)] Yingjun Zhu,[1, a)] Binbin Hong,[2] Xinlin Wu,[3, 4, 5] Jingwen Zhang,[3, 4, 5] Mufeng Cai,[3, 4, 5] Da Zhou,[1, b)] and Yu Liu[3, 4, b)]

[1)] *School of Mathematical Sciences, Xiamen University, Xiamen 361005, China.*

[2)] *Department of Physics, Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China.*

[3)] *Department of Systems Science, Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China.*

[4)] *International Academic Center of Complex Systems, Beijing Normal University, Zhuhai 519087, China.*

[5)] *School of Systems Science, Beijing Normal University, Beijing 100875, China.*

[a)] These authors contribute equally.

[b)] Corresponding authors:

zhouda@xmu.edu.cn (D.Z.), yu.ernest.liu@bnu.edu.cn (Y.L.)

SUPPLEMENTARY NOTES

1. Accurate rate vs. coarse-graining degree

If a connection weight w is within the range $(n - 1/2) \cdot d \leqslant w < (n + 1/2) \cdot d$ (where n is an integer, and d is the coarse-graining interval), we categorize the connection as type n, denoted by a specific symbol. Figure 1 illustrates the variation in accuracy (namely, the change in neural network's performance) with different coarse-graining intervals. It is evident that when d is less than or equal to 0.1, the network's performance is largely maintained before and after coarse-graining. Therefore, in this study, we have chosen $d = 0.1$ for coarse-graining the neural network, enabling its transformation into a set of sequences.

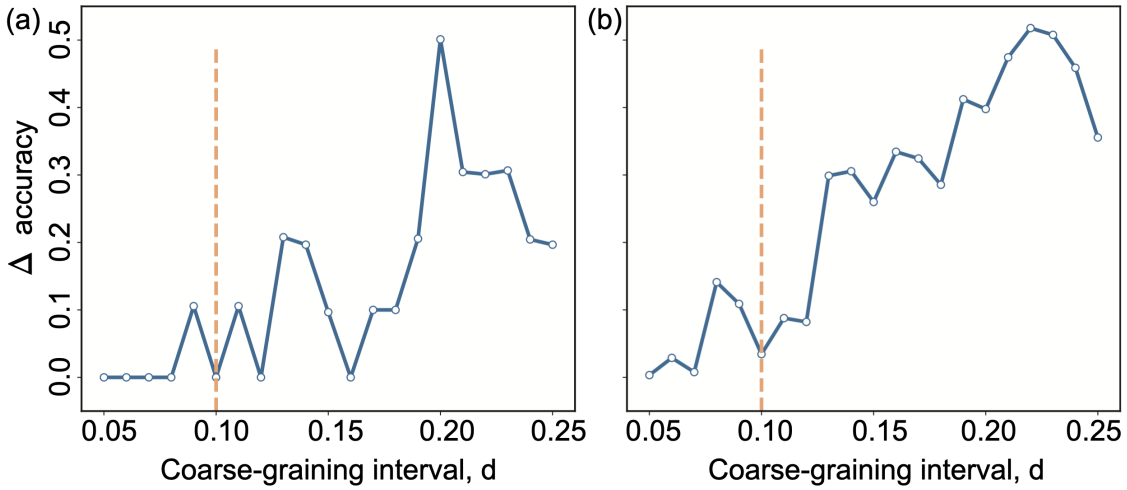

FIG. 1. Variation in accuracy (y-axis) with different coarse-graining intervals. This figure shows the change in accuracy after coarse-graining, calculated as the accuracy of the coarse-grained network minus that of the original network. A value of 0 on the y-axis implies no change in accuracy, while larger values indicate a greater decrease in accuracy, reflecting a reduction in network performance. (a) This is for architecture [3,4,20,2]. (b) For architecture [3,10,13,2].

2. Additional tasks using MLPs

In addition to the odd-even recognition task in the main text, we conducted additional tasks: one more classification tasks, two regression tasks, and two time series forecasting tasks using MLPs, in order to verify our conclusions on a broader level. The results of these five tasks align well with our expectations and corroborate the conclusions drawn in the main text. Please see the detailed descriptions of these five tasks below.

Classification tasks

Besides the odd-even classification task discussed in the main text, we conducted one additional classification task: the famous MONK's Problem[1], which was used to test a wide range of induction algorithms. The dataset consists of six features and outputs two categories. The MLP we utilized randomly selected an architecture for each training session. The constraints on the network architecture are as follows: a network with a single hidden layer is limited to a maximum of 60 neurons per layer, a network with two hidden layers is restricted to a maximum of 20 neurons per layer, and a network with three hidden layers can have up to 15 neurons per layer.

Figure 2 depicts the relationship between the order-rate, η, and the accuracy across 206 different architectures we randomly selected after the same training duration (150 epochs). It was observed that networks with good performance predominantly had η values ranging approximately between 0.4 and 0.6.

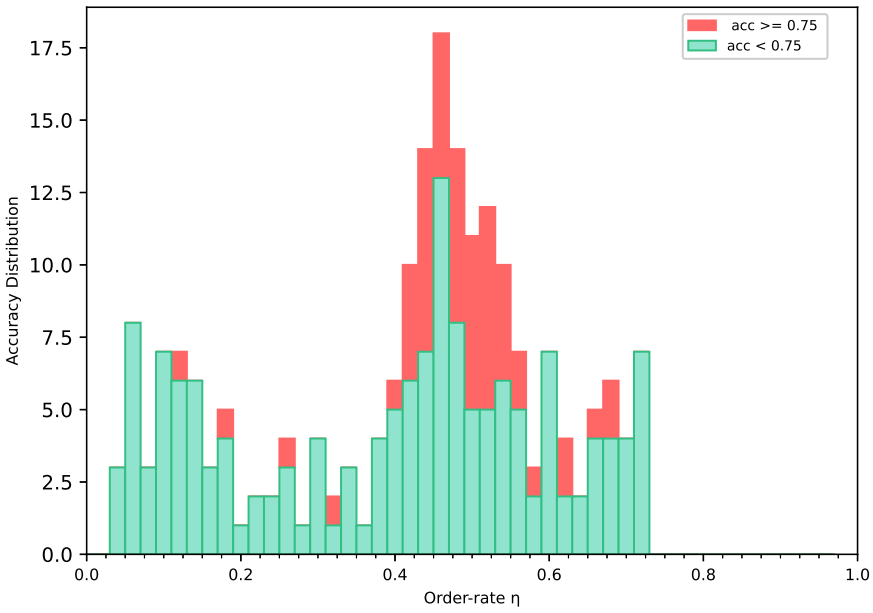

FIG. 2. The distribution of the order-rate η vs. accuracy in the classification task (MONK's Problem). The statistics include all of the aforementioned 206 different architectures.

Regression tasks

Our first regression task focused on concrete compressive strength[2], a highly nonlinear function influenced by age and ingredients. The dataset comprises eight quantitative input variables and one quantitative output variable. The MLP we utilized randomly selects an architecture for each training session. The constraints on the network architecture are as follows: a network with a single hidden layer is limited to a maximum of 60 neurons per layer, a network with two hidden layers is restricted to a maximum of 20 neurons per layer, and a network with three hidden layers can have up to 15 neurons per layer.

For regression tasks, the quality of the model can be evaluated using the mean-square error (MSE). Figure 3 illustrates the relationship between the order-rate, η, and MSE for 565 different architectures we randomly selected after undergoing the same training period (2000 epochs). It was observed that the distribution of η values was relatively concentrated, with networks exhibiting good performance primarily having η values ranging between approximately 0.3 and 0.55.

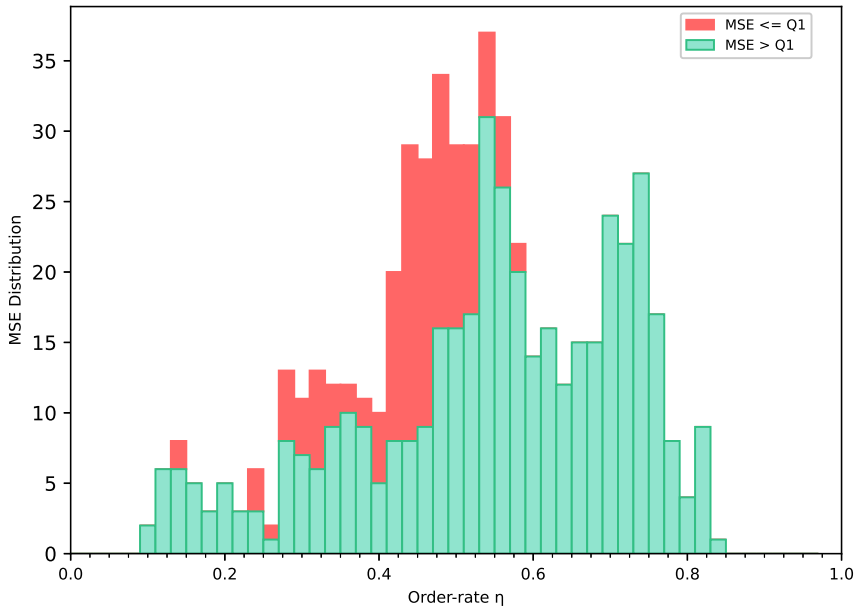

FIG. 3. The distribution of the order-rate η vs. MSE in the regression task (concrete compressive strength). "MSE <= Q1" means that the MSE is less than or equal to the first quartile (Q1). The statistics include all of the aforementioned 565 different architectures.

Our second regression task was centered on city-cycle fuel consumption[3]. The dataset consists of seven quantitative input variables and one quantitative output variable. The selection strategy for the MLP is the same as in the previous regression task. Figure 4 depicts the relationship between the order-rate, η, and MSE across 579 different architectures we randomly selected after the same training duration (2000 epochs). It was observed that networks with good performance predominantly had η values ranging approximately between 0.35 and 0.5.

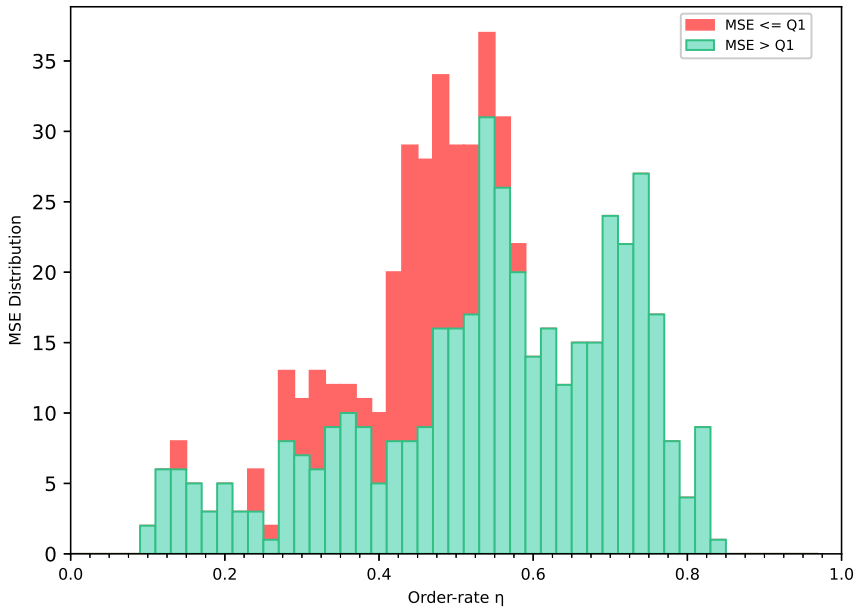

FIG. 4. The distribution of the order-rate η vs. MSE in the regression task (city-cycle fuel consumption). The statistics include all of the aforementioned 579 different architectures.

Time series forecasting tasks

Our initial endeavor in time series forecasting focused on climate prediction. The data was sourced from the meteorological station at the Max Planck Institute for Biogeochemistry in Jena, Germany, and spans the years 2009 to 2016[4]. We utilized the atmospheric pressure data from the previous four days to predict the atmospheric pressure on the fifth day. We structured MLPs with the following specifications: an input layer containing four neurons, a first hidden layer with neuron counts ranging from 2 to 18, and a second hidden layer with 1 to 17 neurons, with a final output neuron. A total of 289 different network architectures were explored under these constraints.

In time series forecasting tasks, the quality of the model is also assessed using MSE. Figure 5 illustrates the relationship between the order-rate, η, and MSE across 289 different architectures following an identical training duration of 200 epochs. It was observed that the distribution of η values was relatively concentrated, with networks exhibiting good performance primarily having η values ranging between approximately 0.45 and 0.65.

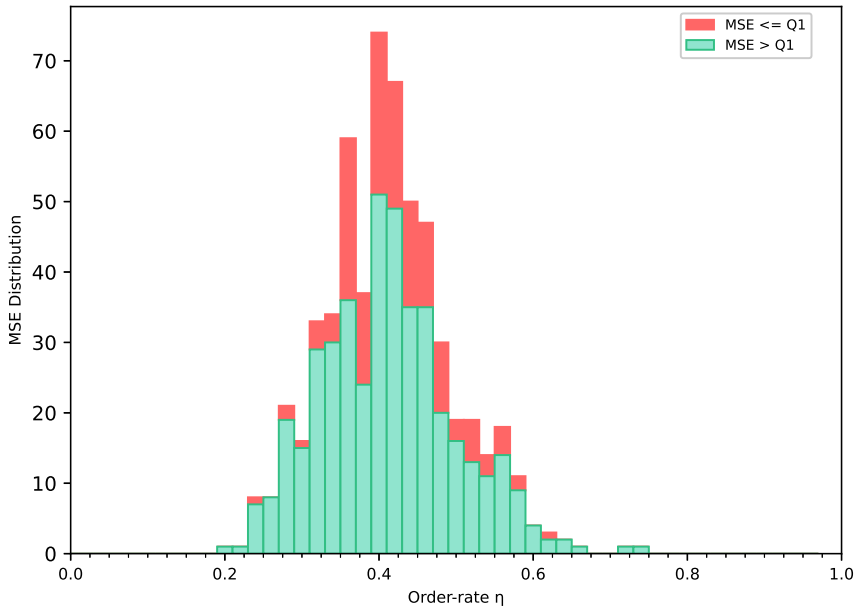

FIG. 5. The distribution of the order-rate η vs. MSE in the time series forecasting task (climate forecasting). "MSE <= Q1" means that the MSE is less than or equal to the first quartile (Q1). The statistics include all of the aforementioned 289 different architectures.

Our second task in time series forecasting focused on the number of air passengers. The dataset includes the count of air passengers from the years 1949 to 1960[5]. We utilized the passenger count from the previous seven months to predict the passenger count on the following month. We designed MLPs with the following configuration: an input layer containing seven neurons, a first hidden layer with neuron counts ranging from 10 to 29, and a second hidden layer with 5 to 24 neurons, culminating in a final output neuron. Under these conditions, a total of 400 different network architectures were explored. Figure 6 depicts the relationship between the order-rate, η, and MSE across 400 different architectures after the same training duration (500 epochs). It was observed that networks with good performance predominantly had η values ranging approximately between 0.45 and 0.55.

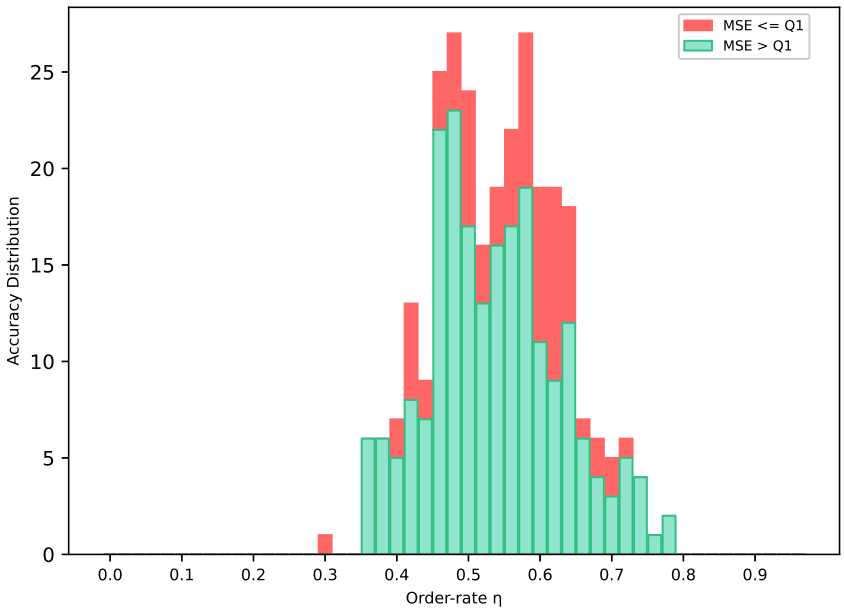

FIG. 6. The distribution of the order-rate η vs. MSE in the time series forecasting task (the number of airline passengers). The statistics include all of the aforementioned 400 different architectures.

SUPPLEMENTARY REFERENCES

[1]J. Wnek. MONK's Problems. UCI Machine Learning Repository, 1992. DOI: https://doi.org/10.24432/C5R30R.

[2]I-Cheng Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C5PK67.

[3]R. Quinlan. Auto MPG. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

[4]Jena Climate 2009-2016. kaggle, 2018. https://www.kaggle.com/datasets/stytch16/jena-climate-2009-2016/data.

[5]Air Passenger Data for Time Series Analysis. kaggle, 2021. https://www.kaggle.com/datasets/ashfakyeafi/air-passenger-data-for-time-series-analysis/data.